Predicting Alzheimer's Disease Progression by Analyzing the Impact of Demographic and Health Factors

By

Manvi Sharma

A Project Submitted to The Department of Public Health Sciences

Biological Sciences Division

The University of Chicago

In Fulfillment of the Requirements for the

Master of Public Health (MPH) Degree

December 2024

1. Introduction

Alzheimer's Disease (AD) is the most prevalent neurodegenerative disorder that causes loss of intellectual and social skills. AD manifests as a progressive, degenerative disorder that attacks nerve cells in the brain, or neurons, resulting in loss of memory, thinking and language skills, as well as behavioral changes[1]. Currently AD is an irreversible process with no cure, affecting approximately 1 in 10 people older than 65 and nearly a third of the population older than 85, significantly impacting the quality of life of aging populations[2].

The social, personal and economic impact of Alzheimer's is profound: AD is currently ranked as the seventh leading cause of death in the United States and is the most common cause of dementia among older adults[3]. The burden of Alzheimer's disease is estimated using disability-adjusted life years (DALYs), which represent years of healthy life lost. In 2019, Alzheimer's disease and related dementias contributed to 33.1 million DALYs lost, making it one of the leading causes of mortality and morbidity among older adults[4]. As the global population continues to age, the burden of the disease is projected to increase. Experts suggest that with over 6 million Americans currently living with the disease[4], the number could potentially triple by 2050. The rise in Alzheimer's prevalence makes its implications in public health very important. Especially when considering the age demographic Alzheimer's majorly affects, and the increased cost of caregiving demands, economic strain and overall societal challenges that are faced by families due to this disease.

Recent advancements have made significant progress in understanding the disease, however, the cause remains a probable combination of many factors including age-related changes in the brain, genetics, environmental and lifestyle factors[5]. Treatments for AD only try and delay the onset of further cognitive decline, but there is no complete cure. The stages of the disease also vary, as no two individuals experience Alzheimer's the same way[6]. Being a

progressive disease, AD does worsen over time at different speeds. The variability in how the disease progresses could be influenced by demographic factors (age, gender, race, ethnicity) and health-related factors (comorbidities, lifestyle, genetics).

Thus, this project aims to provide a predictive model that can identify the *rate of AD disease progression* influenced by patient demographic and health factors. This is achieved through analyzing statistical associations between demographic factors, health factors and Alzheimer's disease progression. The model could potentially help in earlier detection of AD, better resource allocation, improved intervention, and improved quality of life for patients and families. This predictive model could potentially identify individuals or groups at greater risk of disease when finetuned further, and this could aid in addressing health inequities across different communities. Through developing and evaluating this predictive model, the goal of the project is to contribute to public health efforts in mitigating the impact of this disease.

To achieve this aim, the model's performance was strategically evaluated using demographic and health-related features that are available in the dataset obtained from the National Alzheimer's Coordinating Center **(**NACC). The machine learning approach used in this paper focused on metrics derived from longitudinal clinical data, assessing progression trends over multiple patient visits.

The central research questions and hypothesis of the paper include:

- *How demographic and health factors influence Alzheimer's progression as measured by changes in AD progression over time?* It was hypothesized that the factors would significantly predict AD progression.

- *Which variables are the most significant predictors of AD progression?* Based on research, it was hypothesized that physical and mental health indicators such as BMI, anxiety, depression would be the most significant predictors.
- *How accurately can ML models predict the rate of progression using this longitudinal data?* It was hypothesized that this model would have moderate predictive accuracy in estimating the rate of progression based on factors selected, and the exclusion of genetics and biomarker data.

2. Material and Methods

2.1. Data

a. Data Access

For this study, the data used was obtained from the National Alzheimer's Coordinating Center (NACC): The Uniform Data Set (UDS), Version 3. Access to this dataset required an application process, where I described the objectives and implications of the research project. The NACC then provided a tailored dataset with longitudinal clinical, demographic, and health information data specifically related to Alzheimer's patients.

b. Data Structure

The NACC Uniform Data Set (UDS) is an ongoing longitudinal study established in 2005, which tracks participants across multiple visits. The dataset collects data from participants at over 40 Alzheimer's Disease Research Centers (ADRCs) across the United States.

Each row in the dataset corresponds to a single visit for a specific participant, with a unique ID linking multiple records for the same individual. The data collected includes key variables including demographics, health history, cognitive assessments, clinical diagnoses and others. The dataset is continuously updated with new participants and additional longitudinal

data from existing participants. For this study, we used subject demographic features (age, sex, race, education), health indicators (BMI, hypertension, depression), and cognitive measures (Clinical Dementia Rating) [see Table 1]. These predictors provided insights for building a model to estimate Alzheimer's progression based on demographic and health factors.

c. Key Variables

The key variables selected from the NACC UDS that aligned with our research objective of predicting Alzheimer's disease progression. They were categorized in groups: [see Table 1]

1. Subject demographic Variables
2. Health Indicators
3. Cognitive Measures (CDR score)

The Clinical Dementia Rating global score (CDRGLOB) was used as the primary marker for Alzheimer's progression. Time-based variables such as NACCFDYS (days since the initial visit) allowed us to track longitudinal changes.

The global Clinical Dementia Rating (CDR) is a summary score used to stage the severity of dementia based the evaluation of various cognitive domains like memory, orientation, judgment, and daily functioning.[7] CDR provides a single categorical score to indicate the level of dementia a patient is experiencing based on a scale of 0–3: [8]

no dementia	CDR = 0
questionable dementia	CDR = 0.5
mild cognitive impairment	CDR = 1
moderate cognitive impairment	CDR = 2
severe cognitive impairment	CDR = 3

[See Appendix Figure 1. for more details] [8]

The average rate of progression for Alzheimer's was a derived to use as our *target* variable, it was calculated as the change in CDRGLOB over time. Time based change from previous visit was incorporated to capture the dynamics of AD progression.

2.2. Data Preprocessing

The dataset provided by NACC was very large with more than 195,000 rows of patient data. It was preprocessed to ensure the data can efficiently be used for longitudinal analysis and predictive modeling. This process involved handling missing data, creating derived variables, and aggregating records at individual patient level. These steps were essential for capturing trends in Alzheimer's progression.

a. Handling Missing Values

This was done by dropping columns that contained more than 50% null values. Resulting in 126 columns being removed from the dataset. The decision balanced retaining valuable data points while minimizing noise from incomplete records. The columns were identified by calculating the percentage of missing values for each column and then doing an explorative analysis to understand the pattern of missingness. Once the columns were identified as inconsequential data points the model, they were dropped.

The remaining null values in the data were replaced with the mean of the column for numeric columns. And for categorical variables, the placeholder value "Not Found" was added.

b. Filtering for Longitudinal Data

The dataset was then filtered for subjects with multiple visits. To analyze Alzheimer's progression over time, only participants with more than 5 recorded visits were included. This filtering ensured that the dataset focused on individuals with sufficient longitudinal data to be able to analyze progression over time.

c. Derived Variables: Defining Progression Rate

Progression Rate was calculated for each visit as the change in CDRGLOB score divided by the number of days since the previous change in CDR score, this was done for every patient ID. For visits where there was no change in CDRGLOB, the progression rate was 0.

Defining Alzheimer's progression in the model as:

$$Progression\_Rate = \frac{\Delta \text{CDRGLOB}}{\#Days\ since\ change\ in\ score}$$

Lagged versions of the CDRGLOB and NACCFDYS (days from initial visit to each follow-up visit) were created, to calculate differences between consecutive visits.

d. Aggregate Data at Patient Level

The final dataset was created to have patient-level values for every patient ID. This was done by aggregating the longitudinal dataset and grouping by each ID. Producing one average progression rate for each patient over the course of their multiple visits. The aggregated 'Average Progression Rate' was calculated per patient ID to summarize the overall progression trend over the entire observation period for a patient. And was defined as:

$$avg\_progression\_rate = \frac{\Sigma\ \Delta \text{CDRGLOB}}{Total\ Days\ Between\ First\ and\ Last\ Visit}$$

Non-time varying variables like education level obtained, and race were retained as constants, reflecting the data from patients' first visit. Binary variables, like depression and anxiety, were converted into proportions and assigned a True or False value based on thresholds of ≥ 50%.

The resulting final dataset contained a single row per participant with 2654 rows, and contained 19 columns summarizing key demographic, health, and cognitive variables, along with the calculated average progression metrics.

2.3. Model Procedure

a. Feature selection / Feature Engineering

Features for the prediction model were selected based on their theoretical relevance to Alzheimer's progression and correlations with the target variable.

Table 1. The features selected for the final model.

Variable Description	Variable Name	Data Description
Subject Race	NACCNIHR	Categorical
Education Level	EDUC	Numerical
Subject's Sex	SEX	Categorical
Marital Status	MARISTAT	Categorical
Living situation	NACCLIVS	Categorical
Independence Level	INDEPEND	Categorical
Family Cognitive Impairment History	NACCFAM	Binary
Tobacco Use	TOBAC100	Binary
Hypertension (absent or present diagnosis)	HYPERTEN	Binary
Hypercholesterolemia	HYPERCHO	Binary
BMI Level	NACCBMI	Numerical
Vision Status (use of lenses yes/no)	VISCORR	Binary
Depression in last month	DEPD	Binary
Anxiety Diagnosis in last month	ANX	Binary
Total number of visits by patient	total_visits	Count of entries in NACCVNUM
Total days between first and last visit	total_days_visited	Max – Min value within NACCFDYS
Age at first visit	visit_start_age	NACCAGE first visit value
Age at last visit	visit_end_age	NACCAGE last visit value

To capture non-linear relationships between key variables certain interaction terms were generated. And feature engineering was performed when required, where variables with skewed distributions were logarithmically transformed to smooth the data points and make patterns more interpretable. Lastly, in order to incorporate temporal dynamics, lagged versions of CDRGLOB were created. These versions enabled the identification of cognitive changes and allowed for the computation of short-term progression rates when exploring the data.

b. Model development

Two different machine learning models were selected to predict Alzheimer's progression based on their suitability for longitudinal data and their ability to handle non-linear relationships.

1. Random Forest

Random Forest is a robust ensemble learning algorithm that utilizes multiple decision trees to improve predictive accuracy while mitigating overfitting[9]. It is particularly good at handling high-dimensional datasets with multiple features and provides interpretable feature importance metrics while minimizing variance. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size which refers to the minimum number of data points allowed in a single node before further splitting; the number of trees; and the tree depth or number of features sampled. The hyperparameter tuning was conducted using Optuna to optimize each of these three.

2. XGBoost

Extreme Gradient Boost (XGBoost) is a scalable gradient boosting model known for its efficiency and ability to handle missing data. Its regularization techniques prevent overfitting, making it well-suited for datasets with diverse features. The decision trees are also ensemble learning algorithms similar to random forest, however the boosting helps minimize bias and underfitting.[10] With XGBoost, trees are built parallelly, following a level-wise strategy, scanning across gradient values and using partial sums to evaluate the quality of splits at every possible split in the training set.[11] Optuna was used again to optimize the key parameters for this model including the maximum tree depth, learning rate and subsampling/column sampling.

2.4.Evaluation of Model Performance

The predictive performance of our models was evaluated using the following metrics: Mean Squared Error (MSE), R-squared score, and residual analysis. MSE quantifies the average squared difference between predicted and actual values, giving insight into the overall error magnitude. R-squared measures the proportion of variance in Alzheimer's progression explained

by the model, serving as an indicator of how well the model fits. Residual plots were also analyzed to assess model assumptions and detect potential patterns in prediction errors.

For this project, the focus was on optimizing the R-squared values, aiming to build a model that accurately predicts our target- the rate of Alzheimer's progression. Given the medical relevance of tracking disease progression, importance was placed on minimizing unexplained variance to improve predictive reliability.

The evaluation process involved hyperparameter optimization for both models using Optuna, a framework for automated hyperparameter tuning. The data was also cross validated, where the dataset is split into 5 folds, the 5-fold cross-validation was implemented to be able to get a robust estimate of the model's performance by utilizing all data points for both training and testing across 5 iterations. And to ensure reproducibility, all experiments were conducted on a standardized final dataset derived from the NACC UDS data. The evaluation provided a comprehensive analysis of each model's predictive power while maximizing interpretability.

2.5. Methods Justification

a. Strengths and Weaknesses of Progression Definition

The progression rate of Alzheimer's is calculated for each visit as the change in CDR score divided by the number of days since the previous change in score. The definition incorporates both the magnitude of cognitive decline and time elapsed, since AD progression is dynamic over time, the definition can capture and reflect that. The formula is intuitive to interpret from the dataset and can be calculated consistently across all patients. Making it an ideal variable for a comparative analysis. The definition is also well suited for the NACC dataset which has a longitudinal structure, thus the decision to calculate a rate of change.

However, this approach assumes linearity between cognitive change and time, which does not fully capture the complex, non-linear nature of Alzheimer's progression. Additionally, the formula increases variability, particularly for patients with small changes in scores over short periods, which can result in disproportionately high progression rates.

In the final dataset defining Alzheimer's progression using 'avg_progression_rate' allows to focus on long-term trends, smoothing out variability in short-term changes and providing a stable target variable for predictive modeling. This is done since the model's objective is to predict progression as an aggregate trend, avg_progression_rate, can identify overall progression trends while reducing noise caused by variability.

However, the downside of this approach is it obscures temporal patterns within the disease trajectory that the earlier Progression_Rate - calculated at each visit, could capture. At the same time, since Progression_Rate can be overly sensitive to small changes in CDR or time gaps between visits it was not chosen as the main target variable. Future research could combine these metrics to leverage both insights.

b. Strengths and Weaknesses of Model

To predict Alzheimer's progression, two machine learning models are implemented, and they were each chosen for their ability to handle longitudinal structured data with a combination of categorical and numerical features; as well as the ability to explore non-linear relationships. Each model provides unique insights and benefits, making the combined use valuable. Random Forest offers deeper insights into the relative importance of variables, such as age or education, in influencing progression. This interpretability is essential for informing clinical and public health interventions. XGBoost provides more enhanced predictive performance by capturing intricate patterns in the data, particularly in non-linear relationships between predictors and the target. This model also handles missing data better, ensuring more robustness in predictions.

Through employing both models, the study tries to leverage their distinct advantages to ensure more accurate predictions of progression trends. As well as better insights into the key demographic and health factors contributing to Alzheimer's progression. The dual approach enhances the reliability and generalizability of the findings.

3. Results

3.1. Descriptive Statistics

a. Overview of Patient Demographics and Health Factors

The final dataset cohort consisted of 2,654 patients, with a mean age of 74.3 years at their first visit and 81.9 years at their final visit. Patients completed an average of 7.5 visits (Std Dev = 1.8) over a mean follow-up duration of 2,789 days.

Subject demographic data such as educational attainment ranged widely, with a mean of 15.6 years of education in the cohort of patients. While physical health indicators like BMI had an average value of 25.72, which falls within the healthy-to-overweight range based on common BMI classifications. [See Appendix Table 1.]

The distribution of selected features in the dataset were also analyzed, to understand the predictive power of our variables. Age was normally distributed across patients suggesting a well-balanced sampling across the age range from 40-99. This helped ensure that findings are not overly skewed toward one specific age group of AD patients, adding validity to conclusions about progression rates across ages. Sex was roughly equal distributed between men and women within the cohort of AD patients, which enhanced the generalizability of the findings across genders. While simultaneously reducing the potential biases in the conclusion related to gender-specific differences in Alzheimer's progression.

However, Race distribution in our data was skewed, with most participants being white. The distribution is highlighted in the graph below. And communicates a lack of racial diversity in our dataset, this potentially will have implications in our study's generalizability of progression rates across races. The model will only be robust for studying Alzheimer's progression in white populations, which may reflect the dominant demographic at participating centers. But the findings cannot be generalized to other racial groups due to underrepresentation.

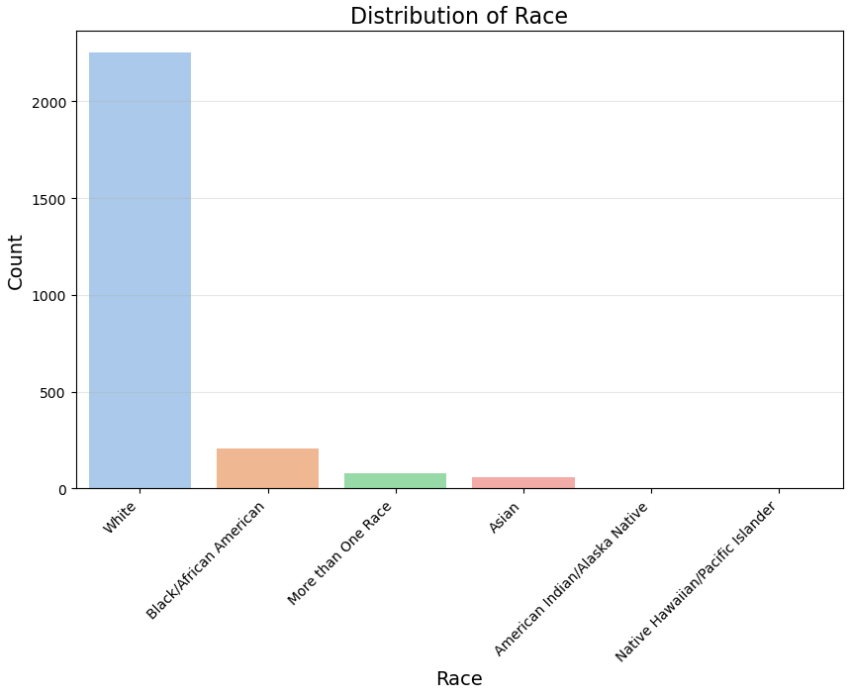


Figure 1. The Distribution Statistics of Race in Dataset.

b. Insights from Progression Rate Plots.

The progression of AD, measured by the average progression rate, was visualized using a spaghetti plot. (see Figure 2. below) The X-axis represents the total number of visits per patient, while the Y-axis denotes the average rate of disease progression. Each line corresponds to an individual patient, capturing the variability in progression rates across the cohort.

A clear trend emerges from the plot, as the number of visits increases, the rate of progression decreases. This suggests that disease progression tends to slow over time, which could be

attributed to factors such as early intervention, treatment effects, or disease-specific dynamics. The substantial heterogeneity among patients remains evident, with some showing consistent progression rates while others show irregular patterns.

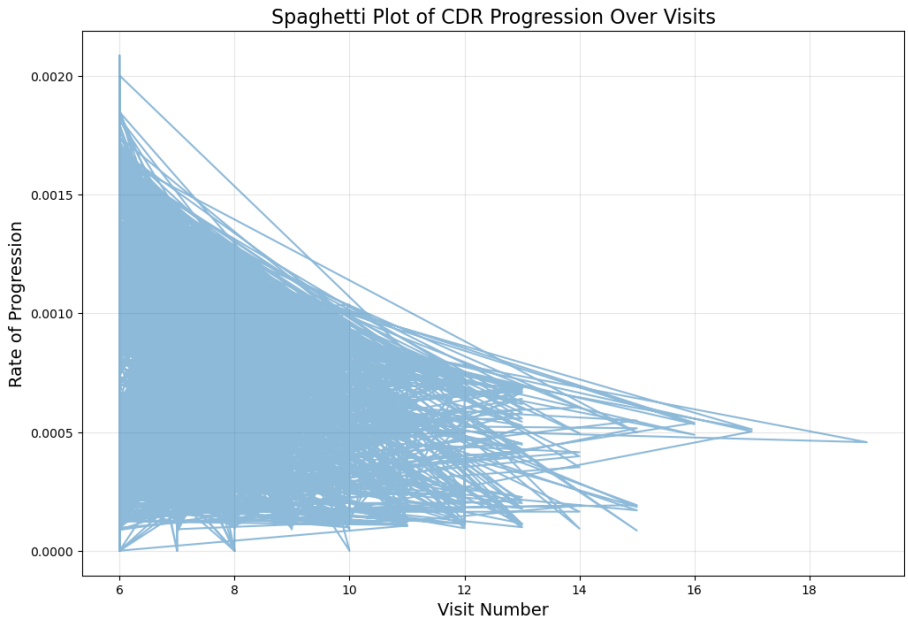


Figure 2. Spaghetti plot of CDR progression rates over number of visits.

The trend of disease progression slowing over time is seen again in the plot below with the 95% confidence intervals (CIs) displayed alongside. The 95% CIs, shaded around the mean CDR line, quantify the variability in progression rates across patients. The widening of CIs at higher visit counts suggests increased variability, perhaps a result of smaller samples of patients with higher number of visit counts.

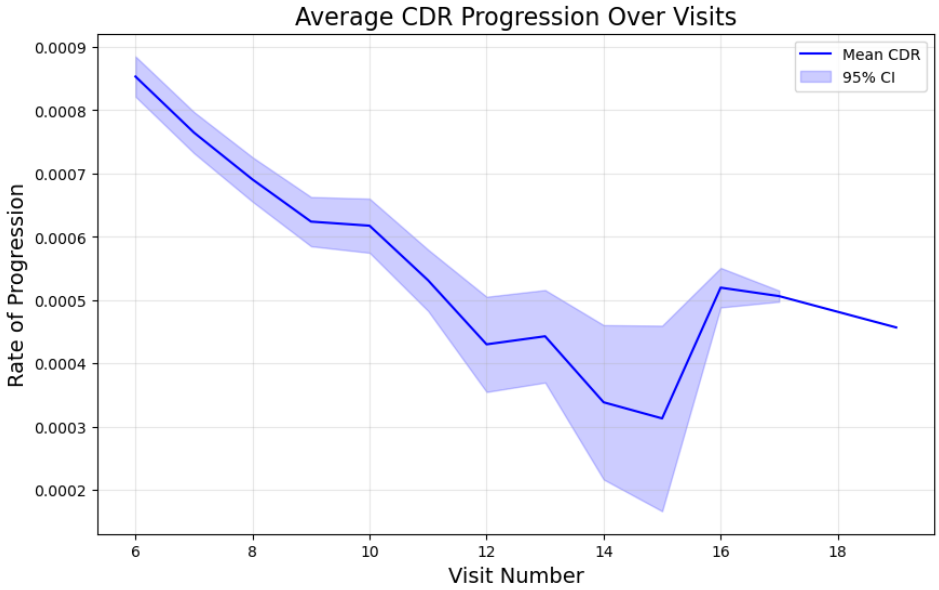


Figure 3. Trend of average CDR progression rate over number of visits.

3.2. Model Performance Evaluation

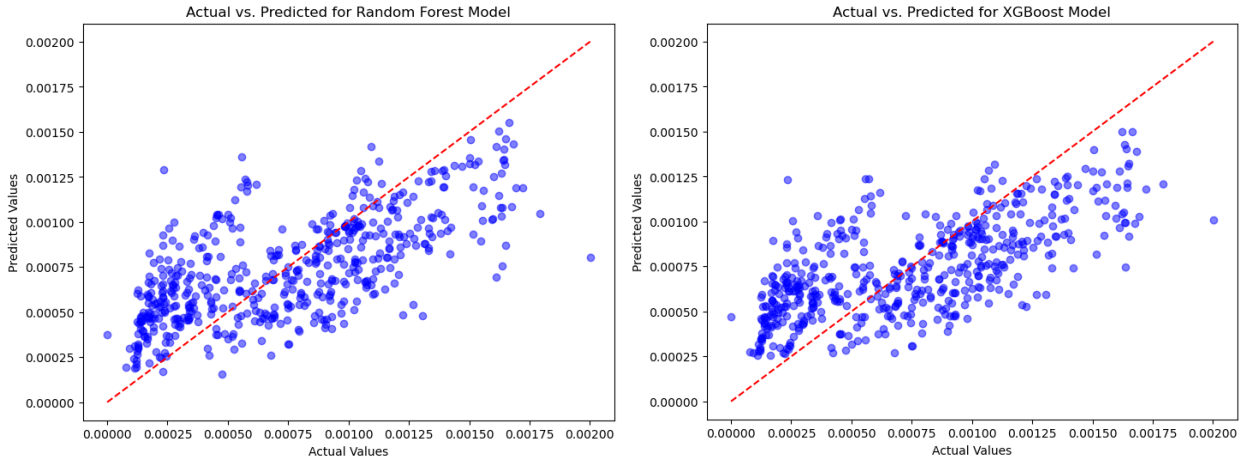
The performance of the Random Forest and XGBoost models was evaluated using Mean Squared Error and R-squared scores displayed below. These metrics were calculated for both the training and test datasets to assess model generalizability and predictive power.

Table 2. The model performance result metrics.

Model	**MSE (test set)**	**R^2 (test set)**
Random Forest	1.05737e-07	0.4595
XGBoost	1.06391e-07	0.4562

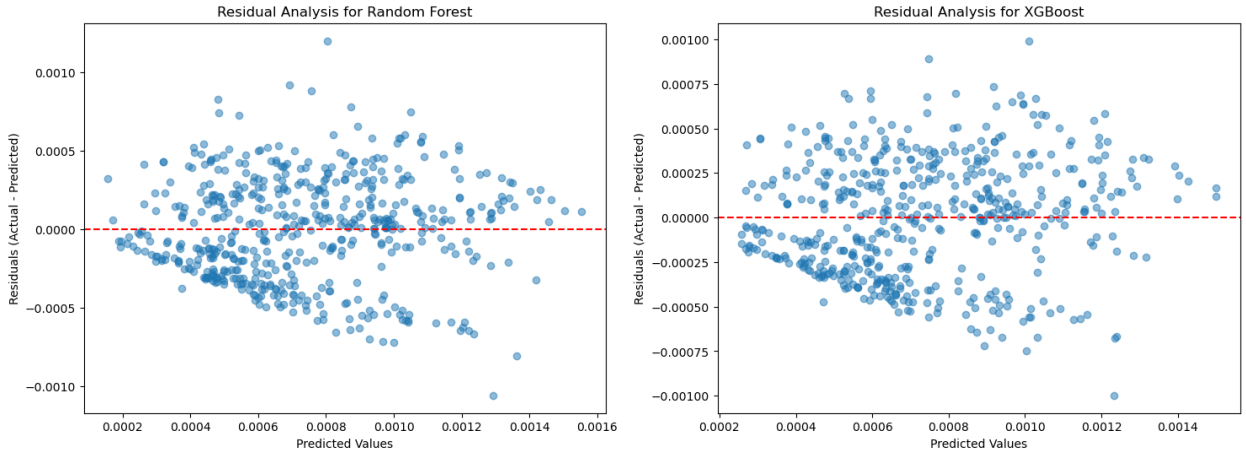
Random Forest and XGBoost both provided very similar performance metrics. The differences between the two models was minimal, indicating comparable performance in predicting Alzheimer's progression rates. Random Forest slightly outperformed XGBoost in terms of both MSE and R^2, suggesting better overall model fit and predictive power for this dataset. However, the difference is too small to conclude one model better than the other.

Figure 4. Actual vs Predicted value plots for both models.



The predicted vs. actual value plots for both Random Forest and XGBoost models (Figure 4.) illustrate the models' predictive accuracy. A strong positive correlation is evident, as most points align along the diagonal line. This indicates that both models generally capture the true progression rates effectively. However, deviations from the diagonal line reflect instances of under or overprediction. The Random Forest model appears slightly more aligned with the diagonal, consistent with its marginally better R-squared value.

Figure 5. Actual vs Predicted value plots for both models.



Residual plots were generated to evaluate the distribution of prediction errors for both models (Figure 5.). Ideally, residuals should be randomly distributed around zero, indicating no

systematic bias in the model's predictions. In the residual plots, both Random Forest and XGBoost show residuals somewhat clustered around zero, suggesting good model performance. However, patterns or spread in residuals may indicate areas where the models could be improved. XGBoost's residuals show slightly more variability, which aligns with its slightly higher prediction error or MSE.

3.3. Feature Importance

Feature importance was evaluated for both Random Forest and XGBoost models to identify the most influential predictors of Alzheimer's disease progression. Feature importance reflects each variable's contribution to improving model performance but does not infer any causality. Feature importance scores, which range from 0 to 1, represent the relative contribution of each feature to the predictive model.

Notably, the highest feature importance values in the analysis of this dataset ranged only between 0.30 and 0.40, suggesting that no single feature overwhelmingly drives the model's predictions. Instead, both the models rely on a combination of features to explain the variance in AD progression rates.

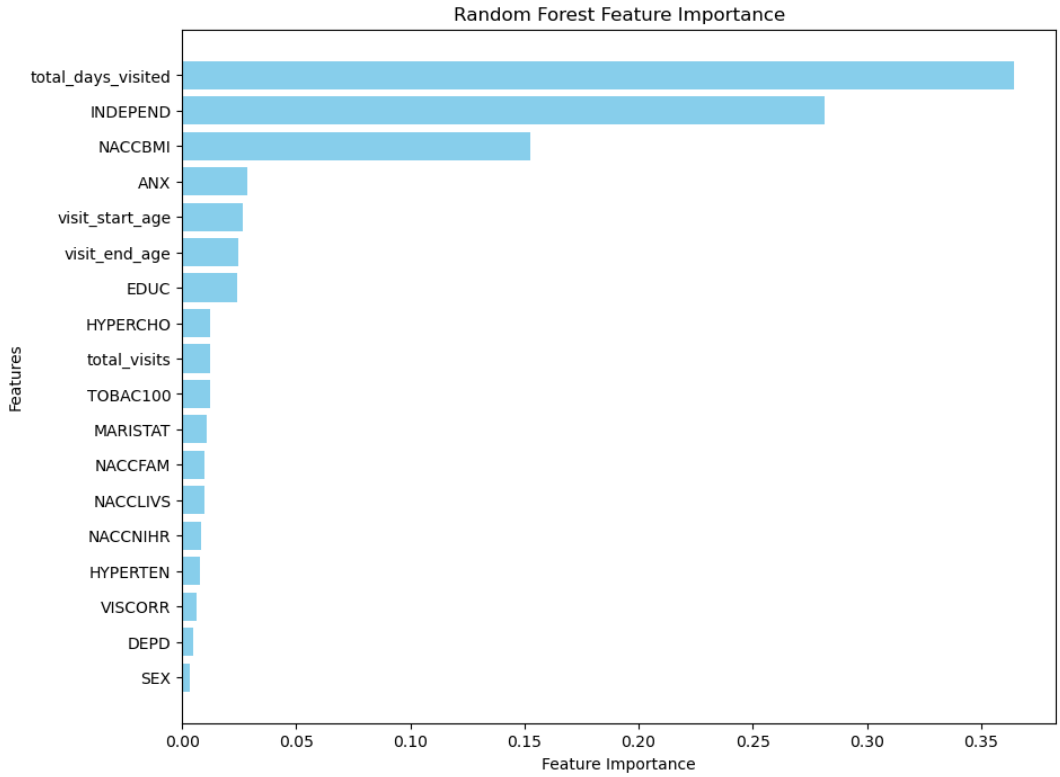


Figure 6. Feature Importance plot for the Random Forest Model.

In the Random Forest model, the top three predictors were Total Days Visited, INDEPEND (measure of functional independence) and NACCBMI (Body Mass Index).

These results suggest that tracking patient independence and physical health over time provides essential insights into disease progression. The strong influence of longitudinal measures like total days visited aligns with clinical findings that Alzheimer's progression is better captured through repeated assessments over time. And the influence of BMI aligns with our hypothesis that physical health characteristics are a significant predisctor for Alzheimer's progression rates.

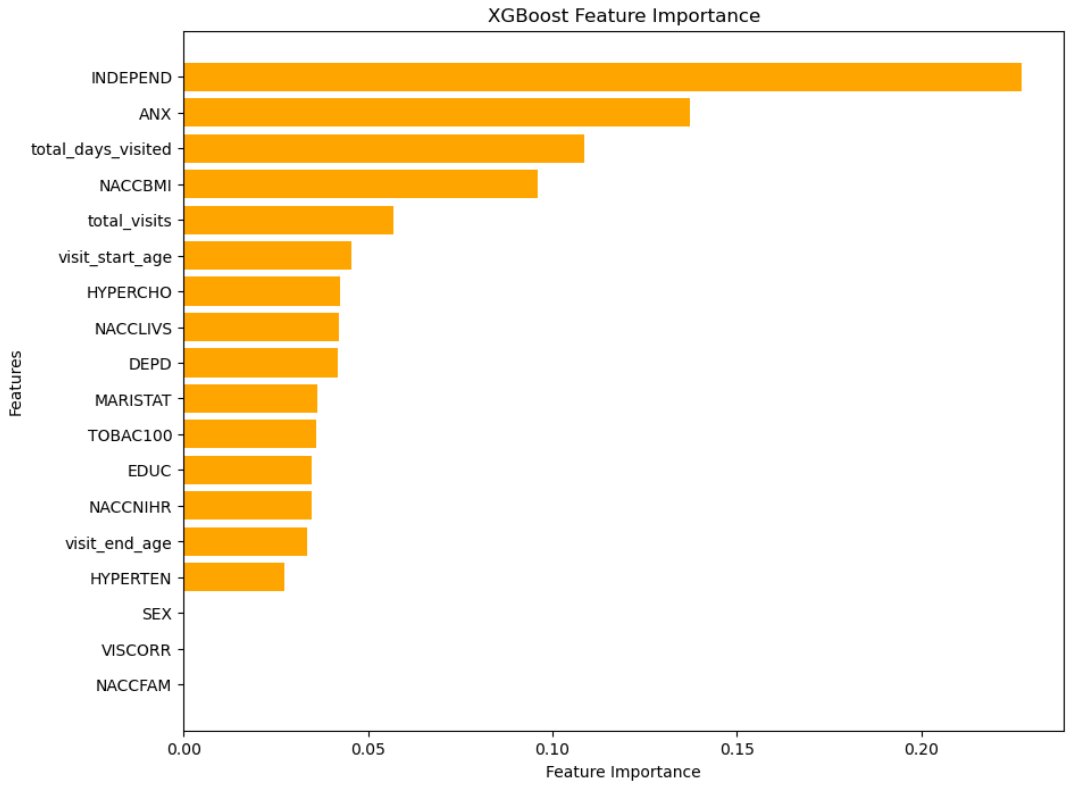


Figure 5. Feature Importance plot for the XGBoost Model.

The XGBoost model identified the same key predictors, suggesting strong consensus regarding their importance in predicting Alzheimer's progression. However, XGBoost highlighted ANX (anxiety) as an additional influential feature, which Random Forest did not emphasize as strongly. This difference likely reflects XGBoost's ability to capture more complex, non-linear relationships within the data. The captured nuanced interactions between mental health and Alzheimer's progression aligns with published evidence that links anxiety to faster cognitive decline and disease progression[12].

The importance of functional independence also aligns with existing literature, which consistently identifies independence as a key marker of disease severity. Loss of independence in early AD is closely related to impaired cognition[13]. Lastly, the significance of BMI reinforces the connection between physical health and cognitive decline, suggesting that maintaining a healthy

BMI and healthy lifestyle may serve as a modifiable risk factor. There is published research that establishes high BMI as a risk factor for AD although the relationship is complex. Obesity in midlife is associated with increased risk for AD, whereas evidence supports both higher and lower than normal BMI increasing risk for AD in late life[14].

While feature importance provides valuable insights, it does not imply causality and comes with its set of limitations. Tree-based models, such as Random Forest and XGBoost, measure the contribution of features to model predictions but may not fully account for confounding variables or complex interdependencies. Additionally, there being differences in feature importance rankings between the two models which have similar performance evaluations metrics – highlights algorithmic structure significantly influences the feature selection too.

4. Discussion

4.1. Public Health Relevance

This research project provides one approach to predicting Alzheimer's disease progression through advanced machine learning techniques applied to a longitudinal dataset from the NACC. The analysis of 2,654 patients revealed certain patterns of disease progression by leveraging demographic and health related features. The prediction model on its own is a small singular dataset, however on a macro-level more robust ML forecasting tools could predict Alzheimer's rapid progression risk at an individual level. Potentially allowing healthcare providers to prioritize interventions in individuals or groups and tailor the healthcare responses. Being able to anticipate patient needs more accurately is particularly more significant in complex and progressive diseases such as Alzheimer's. This approach not only potentially improves patient outcomes for AD but can also assists policymakers in more strategically allocating healthcare resources. The model provides a data-driven framework for understanding how various

demographic and health factors interact to influence disease progression, which could be instrumental in developing more nuanced and effective public health responses to other types of dementia as well.

4.2. Limitations and Future Direction

The findings and methods of this research could present broader implications for public health research and predictive modeling for Alzheimer's and other related dementia disorders. By demonstrating the potential machine learning techniques have in analyzing longitudinal clinical data, this study contributes to the growing area of health care research exploring methods of disease prediction.

Future research could explore how this modeling approach can be applied to other chronic neurodegenerative diseases. However, to further enhance this current model's capabilities, the proposed future next steps can be incorporating additional data sources such as biomarkers and imaging data, as this can provide more robust insights into disease progression and enhance predictive accuracy. Also exploring more sophisticated machine learning techniques, like deep learning algorithms, could also improve predictive accuracy of this project and uncover more relationships within the dataset.

While this project offers some valuable insights, several limitations must be acknowledged. The current model's predictive capabilities are constrained due to the relatively small scale of the target variable. Also, the characteristics of the NACC dataset limit the extent of the analysis. Since there was limited generalizability of the findings, future research should aim to validate these results using more diverse and comprehensive data.

5. Conclusion

Th study utilized longitudinal data from the NACC dataset to develop and evaluate a machine learning model for predicting Alzheimer's disease progression. The research aimed to address three primary research questions, giving insights into the complex dynamics of AD progression.

Demographic and Health Factors Influencing AD Progression

Our first research hypothesis proposed that demographic and health factors would significantly predict AD progression. The findings support this hypothesis to an extent. Both Random Forest and XGBoost models demonstrated that multiple demographic and health variables do contribute to predicting disease progression, however the features importance varies.

Significant Predictors of AD Progression

The second hypothesis speculated that physical and mental health indicators would be the most significant predictors. Our analysis partially confirmed this hypothesis, with some nuanced findings. BMI and anxiety were indeed identified as important predictors, aligning with existing literature. But functional independence emerged as an even more critical predictor, establishing importance of tracking this measure and conducting comprehensive health assessments on AD patients to track their disease progression.

Predictive Accuracy of Machine Learning Models

The third hypothesis anticipated moderate predictive accuracy, given the exclusion of genetic and biomarker data. The results align with this expectation. The R-squared scores of the Random Forest and XGBoost models demonstrated moderate but meaningful predictive capabilities of explaining approximately 46% of the variability.

References

1. Alzheimer's Foundation of America. Alzheimer's Foundation of America [Internet]. Alzheimer's Foundation of America. 2017. Available from: https://alzfdn.org/
2. Cleveland Clinic. Alzheimer's disease [Internet]. Cleveland Clinic. 2022. Available from: https://my.clevelandclinic.org/health/diseases/9164-alzheimers-disease
3. CDC. About Alzheimer's [Internet]. Alzheimer's Disease and Dementia. 2024. Available from: https://www.cdc.gov/alzheimers-dementia/about/alzheimers.html
4. Nandi A, Counts N, Chen S, Seligman B, Tortorice D, Vigo D, et al. Global and Regional Projections of the Economic Burden of Alzheimer's Disease and Related Dementias from 2019 to 2050: A Value of Statistical Life Approach. eClinicalMedicine [Internet]. 2022 Sep 1;51(101580):101580. Available from: https://www.sciencedirect.com/science/article/pii/S2589537022003108
5. National Institute on Aging. Alzheimer's Disease fact sheet [Internet]. National Institute on Aging. National Institutes of Health; 2023. Available from: https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet
6. Alzheimer's Association. What is Alzheimer's Disease? [Internet]. Alzheimer's Disease and Dementia. Alzheimer's Association; 2024. Available from: https://www.alz.org/alzheimers-dementia/what-is-alzheimers
7. Khan TK. Clinical Diagnosis of Alzheimer's Disease. Biomarkers in Alzheimer's Disease. 2016;27–48.
8. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993 Nov 1;43(11):2412–2.
9. IBM. What Is Random Forest? | IBM [Internet]. www.ibm.com. IBM; 2023. Available from: https://www.ibm.com/topics/random-forest
10. Gradient Boosting, Decision Trees and XGBoost with CUDA [Internet]. NVIDIA Technical Blog. 2017. Available from: https://developer.nvidia.com/blog/gradient-boosting-decision-trees-xgboost-cuda/
11. Nvidia. What is XGBoost? [Internet]. NVIDIA Data Science Glossary. 2024. Available from: https://www.nvidia.com/en-us/glossary/xgboost/
12. RSNA. Anxiety Associated with Faster Alzheimer's Disease Onset [Internet]. Rsna.org. 2017 [cited 2024 Dec 8]. Available from: https://press.rsna.org/timssnet/media/pressreleases/14_pr_target.cfm?ID=2231
13. Vidoni ED, Honea RA, Burns JM. Neural Correlates of Impaired Functional Independence in Early Alzheimer's Disease. Journal of Alzheimer's disease : JAD [Internet]. 2010;19(2):517–27. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891926/
14. Moody JN, Valerio KE, Hasselbach AN, Prieto S, Logue MW, Hayes SM, et al. Body Mass Index and Polygenic Risk for Alzheimer's Disease Predict Conversion to Alzheimer's Disease. Le Couteur D, editor. The Journals of Gerontology: Series A. 2021 Apr 21;76(8):1415–22.

Appendix

Table 1. Overview of Patient Demographics and Health Factors: Descriptive statistics for variables:

Variable	Mean	Std Dev	Min	25th Percentile	Median	75th Percentile	Max
Visit Start Age	74.3	9.0	34.0	69.0	75.0	80.0	99.0
Visit End Age	81.9	9.2	42.0	76.0	83.0	88.0	108.0
Total Visits	7.5	1.8				8.0	19.0
Total Days Visited	2789	865.4	1438	2156	2578	3268	6565
Years of Education	15.25	3.4	0.0	12.0	16.0	18.0	25.0
Body Mass Index (BMI)	25.72	5.2	10.0	22.9	25.6	28.7	49.3

Table 2. Integration of Capstone Research Project with MPH Competencies

Competency Name	***Competency Substantiation***
MPH Core Competency: Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate.	The project demonstrates this competency through utilizing multiple machine learning models for predictive modeling, utilizing data preprocessing techniques, and computational programming in Python. I also applied statistical techniques like cross-validation and computational methods to extract insights from longitudinal data.
MPH Data Science Competency: Identify the appropriate type of multivariable statistical model to represent relationships between different types of variables.	The project demonstrates this competency by carefully selecting 2 complementary machine learning models - Random Forest and XGBoost. They were chosen to handle mixed data types within longitudinal data and capture non-linear relationships. I also compared model performance using multiple metrics – MSE and R-squared. This was chosen to evaluate complex interaction in the data.
MPH Data Science Competency: Design data visualizations to interpret and communicate research findings.	The project demonstrates this competency through creating multiple informative visualizations including- spaghetti plot showing AD progression rates, confidence interval plot, actual vs. predicted value plots, residual plots, distribution graphs, feature importance plots. These visualizations were made to

	communicate complex research findings and highlight model performance.

Figure 1. The Clinical Dementia Rating (CDR) Scale

(Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993 Nov 1;43(11):2412–2. https://doi.org/10.1212/wnl.43.11.2412-a)

	Impairment				
	None 0	**Questionable 0.5**	**Mild 1**	**Moderate 2**	**Severe 3**
Memory	No memory loss or slight inconstant forgetfulness	Consistent slight forgetfulness; partial recollection of events; "benign" forgetfulness	Moderate memory loss; more marked for recent events; defect interferes with everyday activities	Severe memory loss; only highly learned material retained; new material rapidly lost	Severe memory loss; only fragments remain
Orientation	Fully oriented	Fully oriented except for slight difficulty with time relationships	Moderate difficulty with time relationships; oriented for place at examination; may have geographic disorientation elsewhere	Severe difficulty with time relationships; usually disoriented to time, often to place	Oriented to person only
Judgment and Problem Solving	Solves everyday problems and handles business and financial affairs well; judgment good in relation to past performance	Slight impairment in solving problems, similarities, and differences	Moderate difficulty in handling problems, similarities, and differences; social judgment usually maintained	Severely impaired in handling problems, similarities, and differences; social judgment usually impaired	Unable to make judgments or solve problems
Community Affairs	Independent function at usual level in job, shopping, and volunteer and social groups	Slight impairment in these activities	Unable to function independently at these activities although may still be engaged in some; appears normal to casual inspection	No pretense of independent function outside home Appears well enough to be taken to functions outside a family home	Appears too ill to be taken to functions outside a family home
Home and Hobbies	Life at home, hobbies, and intellectual interests well maintained	Life at home, hobbies, and intellectual interests slightly impaired	Mild but definite impairment of function at home; more difficult chores abandoned; more complicated hobbies and interests abandoned	Only simple chores preserved; very restricted interests, poorly maintained	No significant function in home
Personal Care	Fully capable of self-care		Needs prompting	Requires assistance in dressing, hygiene, keeping of personal effects	Requires much help with personal care; frequent incontinence

Figure 2. Overview of Patient Demographics and Health Factors: Distribution patterns for Age and Sex in the dataset.